Bioinformatics

CS300

Blast, Substitution Matrices and Protein Alignments

(Chap 4 and 5 in textbook)

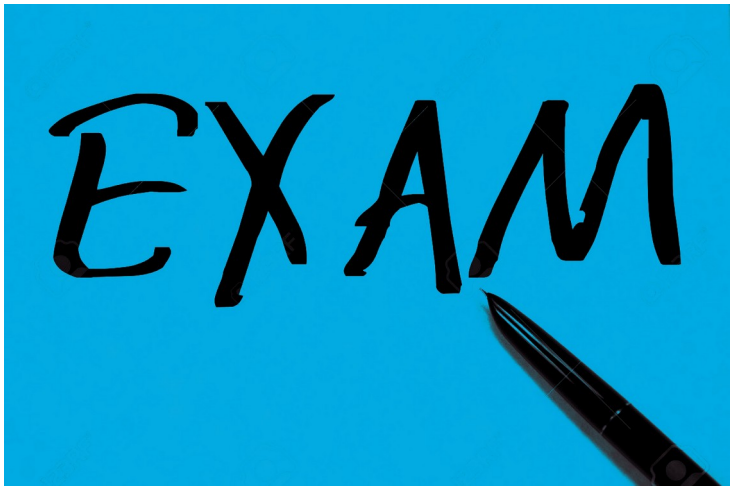
Week8, Deck 2

Fall 2022

Oliver BONHAM-CARTER

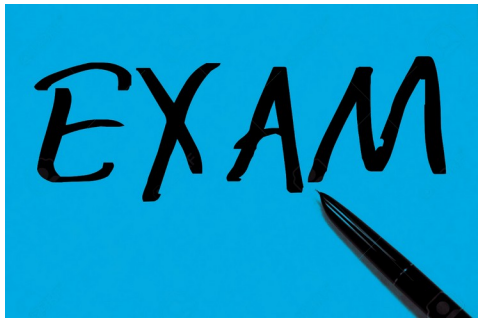
Exam 1
Given out in class, 28th Oct 2022

- Differences between DNA and RNA
- Basic Python programming: syntax, keywords and definitions (covered in class)
- Global and local alignment
- Terms and definitions
- Slides

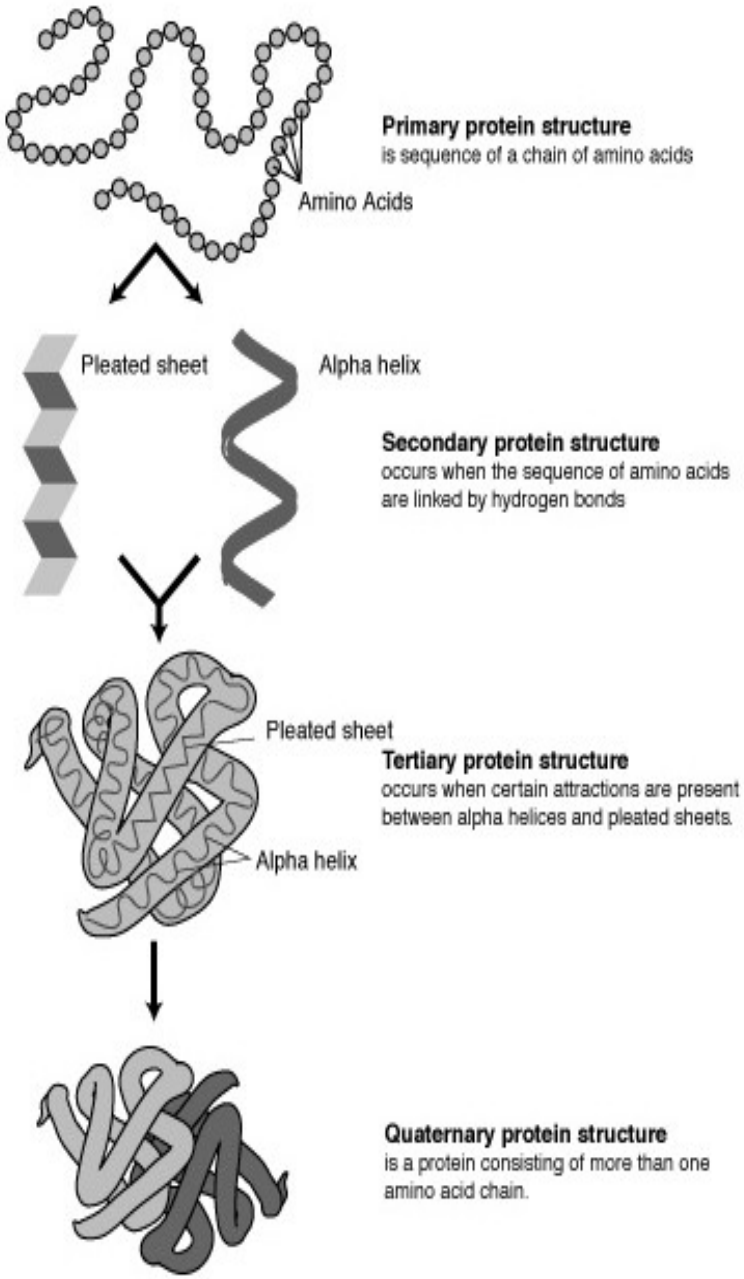


Exam 1
Given out in class, 28th Oct 2022

- Some questions will be involved and will resemble small class activities
  - Central Dogma of Biology
  - Transcription, Translation
  - Detecting genetic disorders
  - Biopython coding and debugging
- Open notes, but Study your notes!



Amino Acids Determine Protein's Shape and Function



The hierarchy of protein structure. Public domain image from The National Genome Research Institute

Similar and Dissimilar Substitution

- Nucleotides – any (base) substitution makes the genetics "different" *in some way*
  - *Substituting similar ones is likely to retain protein structure and function*
  - *Substituting dissimilar ones is likely to change protein structure and disrupt function*

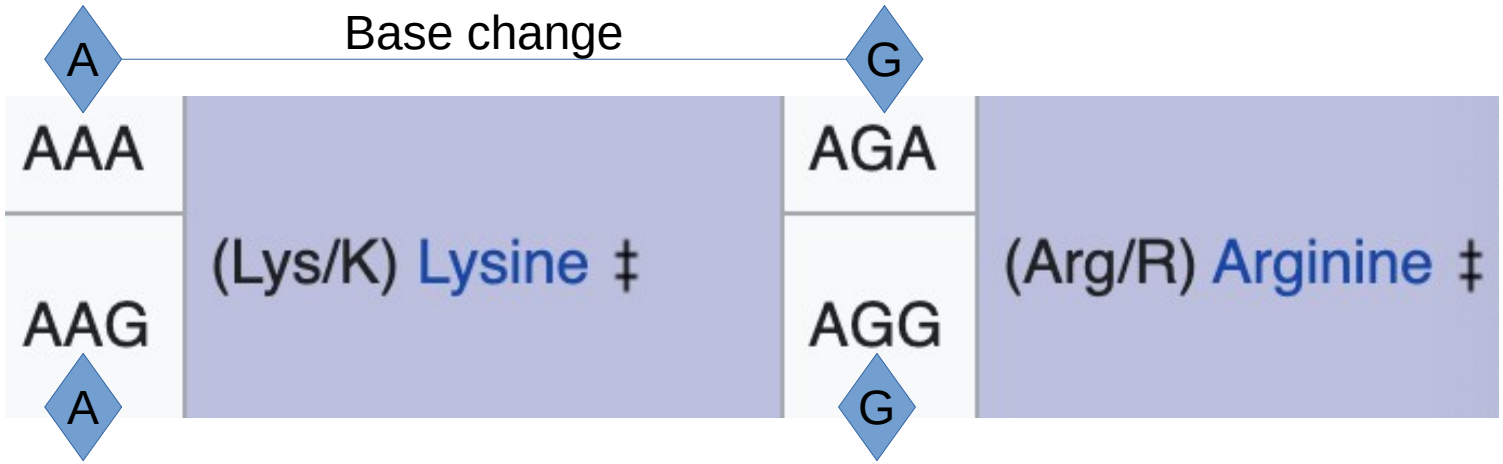
Wait!
How do we spot mutations in protein?

Codon Table

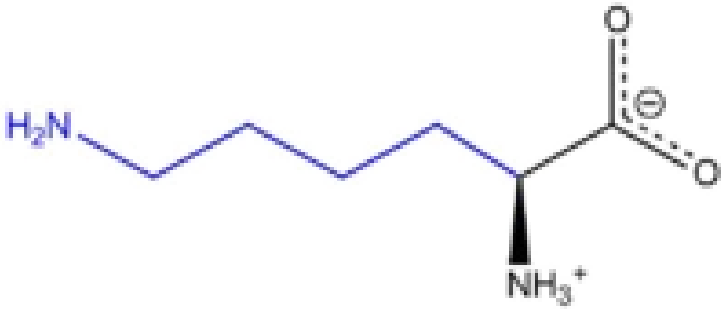
1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine ↑	UCU	(Ser/S) Serine †	UAU	(Tyr/Y) Tyrosine †	UGU	(Cys/C) Cysteine †	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine ↑	UCA		UAA	Stop (*Ochre*) *[note 2]	UGA	Stop (*Opal*) *[note 2]	A
	UUG →		UCG		UAG	Stop (*Amber*) *[note 2]	UGG	(Trp/W) Tryptophan ↑	G
C	CUU		CCU	(Pro/P) Proline ↑	CAU	(His/H) Histidine ‡	CGU	(Arg/R) Arginine ‡	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine ↑	ACU	(Thr/T) Threonine †	AAU	(Asn/N) Asparagine †	AGU	(Ser/S) Serine †	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	AUG →	(Met/M) Methionine ↑	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine ↑	GCU	(Ala/A) Alanine ↑	GAU	(Asp/D) Aspartic acid ↓	GGU	(Gly/G) Glycine ↑	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid ↓	GGA		A
	GUG →		GCG		GAG		GGG		G

Protein: Chemistry

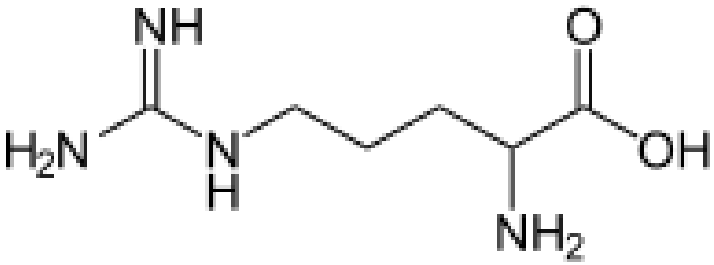
Chemical complexes being replaced by similar chemical complex.
Ex: Arginine (Arg) and Lysine (Lys)
Can this substitution cause harm, now or later?!



Lysine

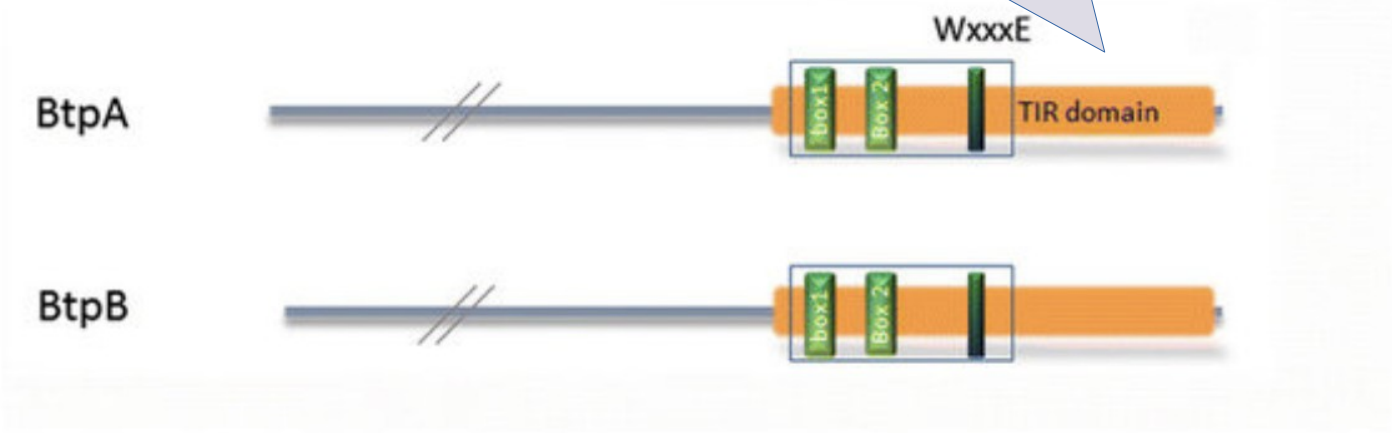


Arginine

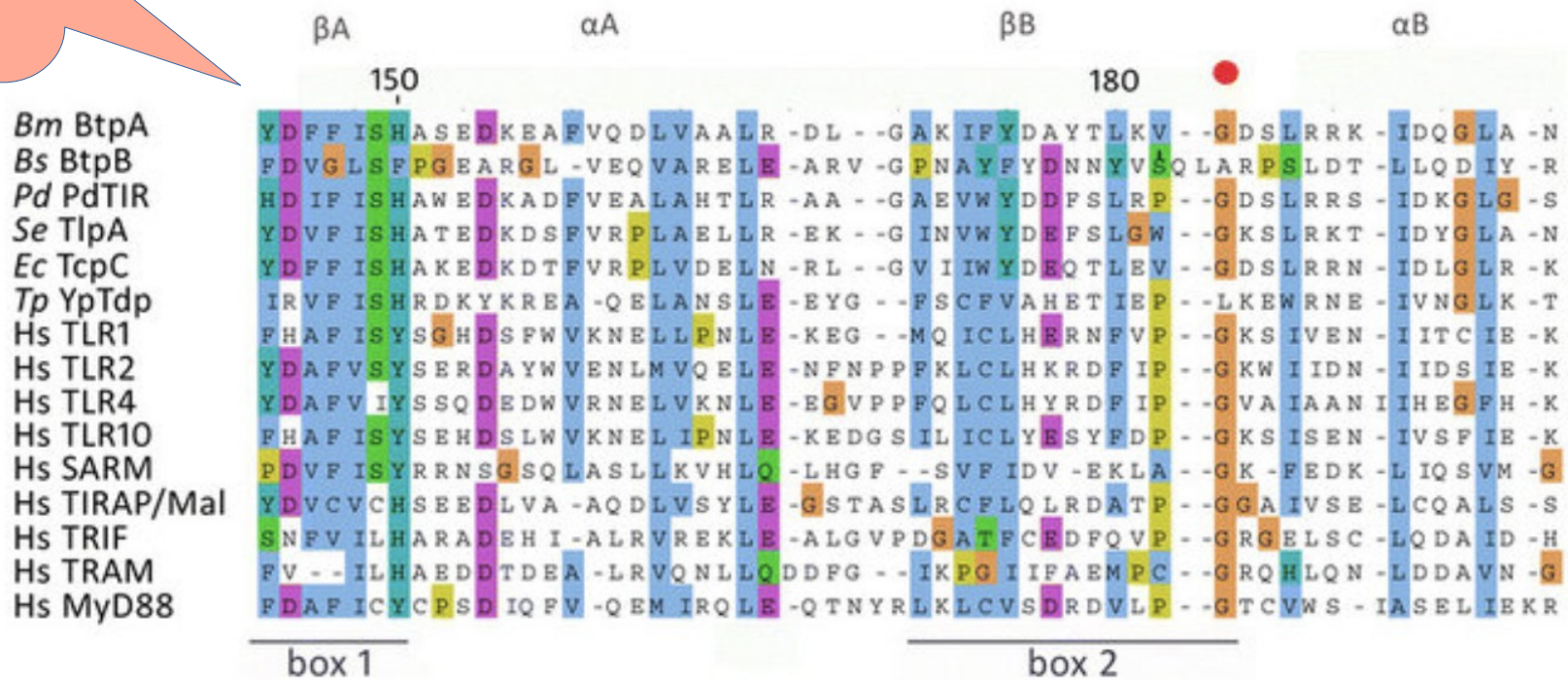


Alignment of Protein *Domains*: the “Functional” Parts of Protein

These domains have individual functions



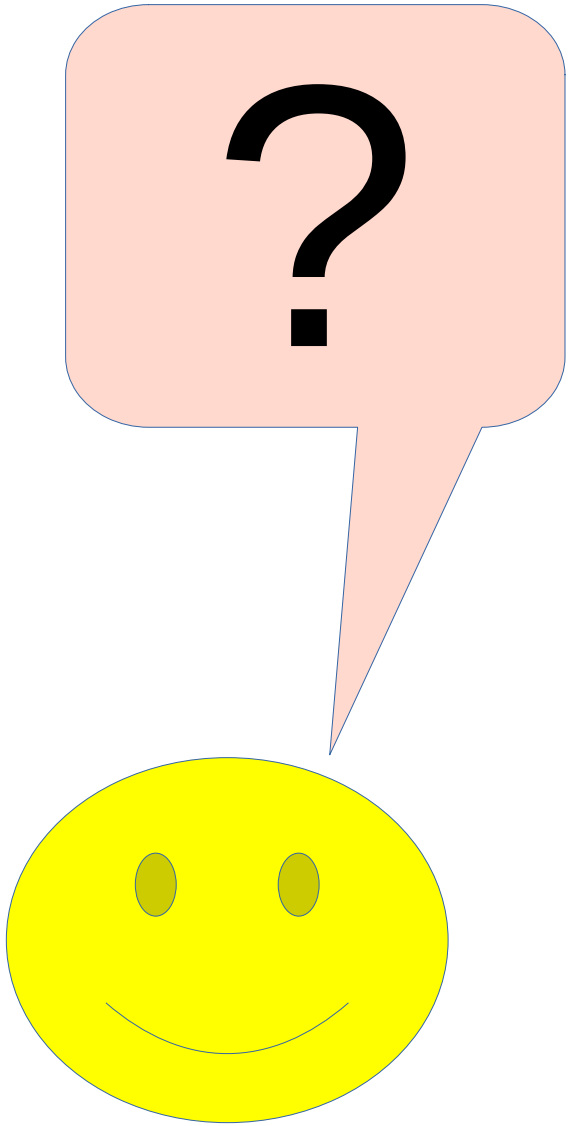
A change in the domain composition may change Function!



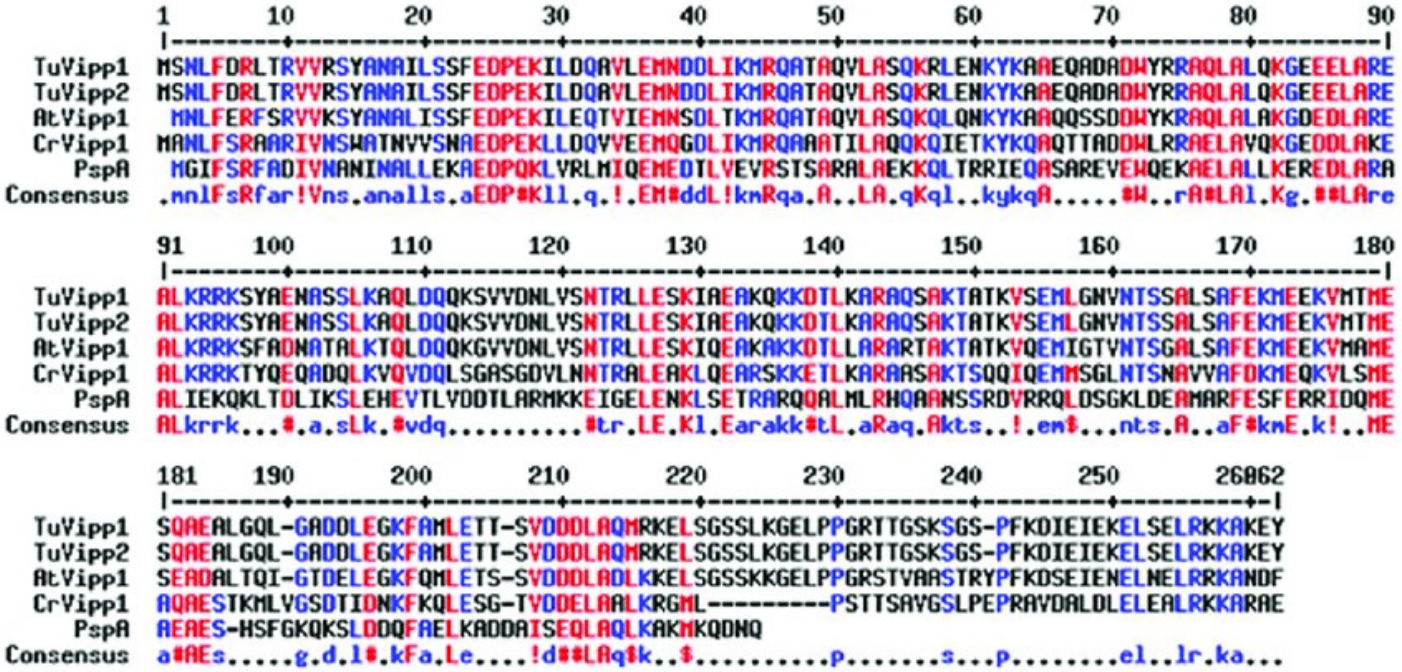
Felix, Christine, et al. "The Brucella TIR domain containing proteins BtpA and BtpB have a structural WxxxE motif important for protection against microtubule depolymerisation."Â Cell Communication and SignalingÂ 12.1 (2014): 1-15.

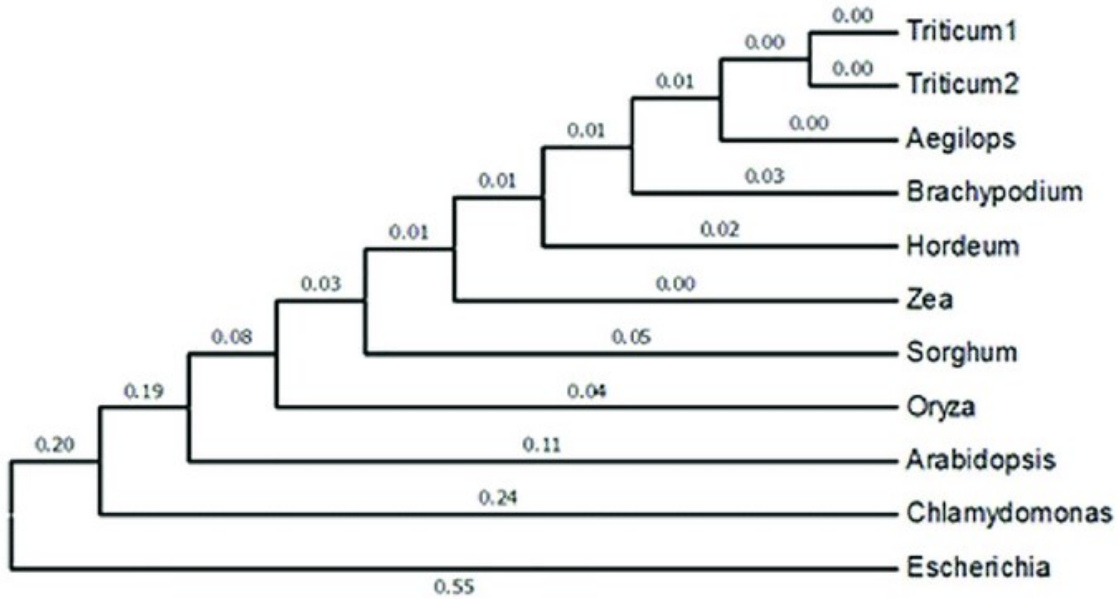
When Comparing Proteins ...

- How are protein sequences different?
- How much difference is there?
- Was this difference due to chance?
- Could the altered protein have a new and different function?
- Did this alteration happen to a domain to impact function?



Blast Also Works With Proteins!!





Gao, Fei, et al. "Two novel vesicle-inducing proteins in plastids 1 genes cloned and characterized in Triticum urartu." PloS one 12.1 (2017): e0170439.

Spotting Differences

A difference: results may not have been experimentally observed, DNA can be translated to produce this protein.

The reading frame of the DNA might produce a different protein than this one

Download GenPept Graphics

PREDICTED: serine/threonine-protein kinase PINK1, mitochondria

Sequence ID: XP_014893419.1 **Length:** 575 **Number of Matches:** 1

Range 1: 1 to 334 GenPept Graphics

Score	Expect	Method	Identities	Positives
653 bits(1684)	0.0	Compositional matrix adjust.	319/334(96%)	328/334

```
Query  1    MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV
            MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV
Sbjct  1    MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV

Query  61   RTSLRGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLEE
            RTSL+GLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLE+
Sbjct  61   RTSLKGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLED

Query  121  VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFAHE
            VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQF+HE
Sbjct  121  VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFSHE

Query  181  DNEVEVQNVRSAACCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI
            DNE EVQNVRS +CCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI
Sbjct  181  DNEEEVQNVRSPSCCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI

Query  241  HITLDGHFGVLPKRVSAHPNVIRVYRAFTADVPLLPGAEEEYPDVLE
             ITLDG FGVLP+RVSAHPNVIRVYRAFTADVPLLPGA+EEYPDVLE
Sbjct  241  QITLDGRFGVLPRRVSAHPNVIRVYRAFTADVPLLPGAQEEYPDVLE

Query  301  LFLVMKNYPYTLRQYLQVSTPNRRQGSLMVLQLL   334
            LFLVMKNYP TLRQYLQVSTPNRRQGSLMVLQLL
Sbjct  301  LFLVMKNYPCTLRQYLQVSTPNRRQGSLMVLQLL   334
```

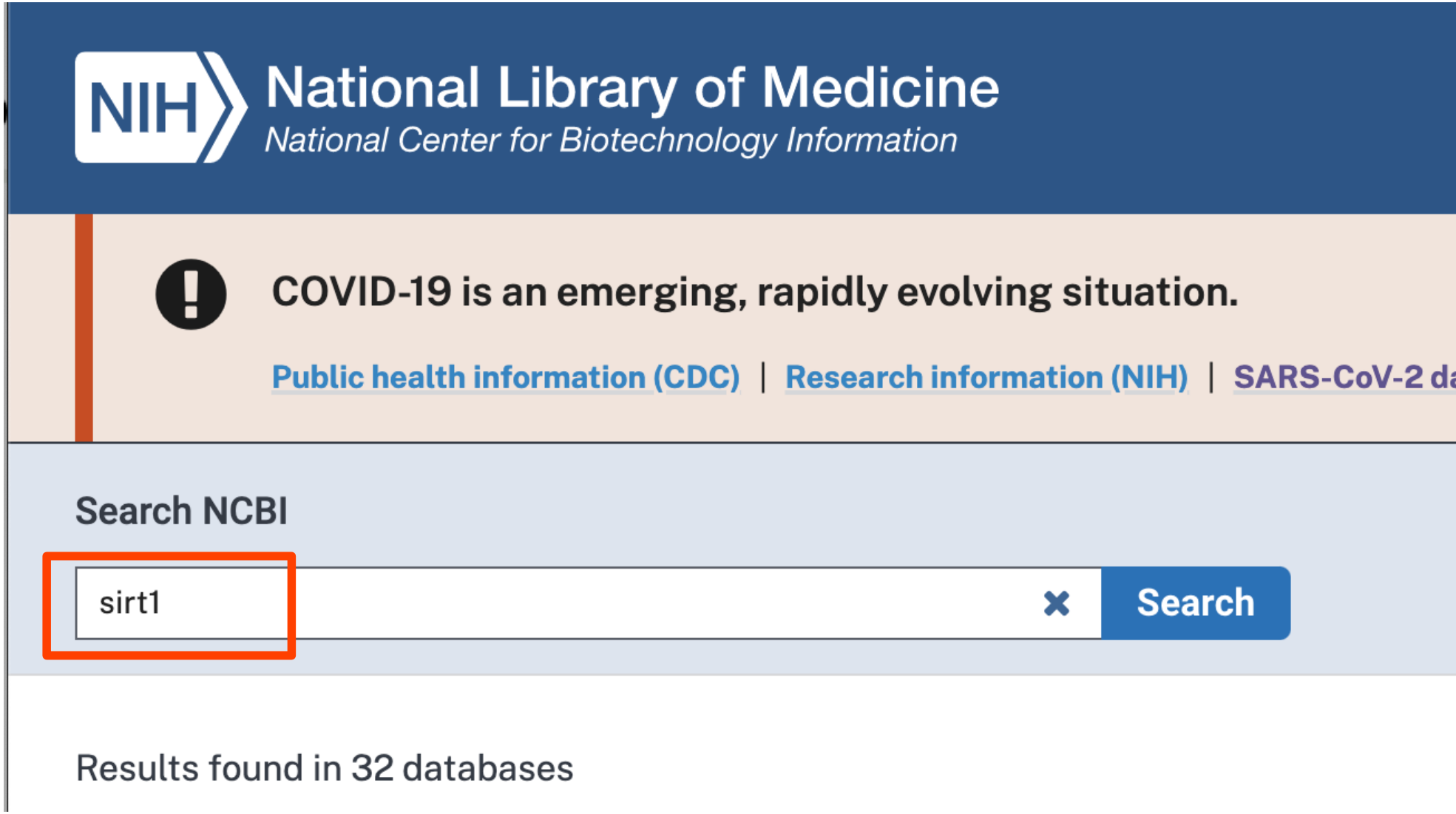
Awesome!

(*But not as awesome as this!*)

Blast-Off!!

- Let’s blast some protein sequences
- https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr_Query_98931

Blasting Proteins



Blasting Proteins

Select Protein

For example, choose this one **CAC5409616** and use Blast link in the record

Proteins

Conserved Domains	9
Identical Protein Groups	570
Protein	6,651
Protein Family Models	45
Structure	129

☐ 2. **SIRT1** [Mytilus coruscus]

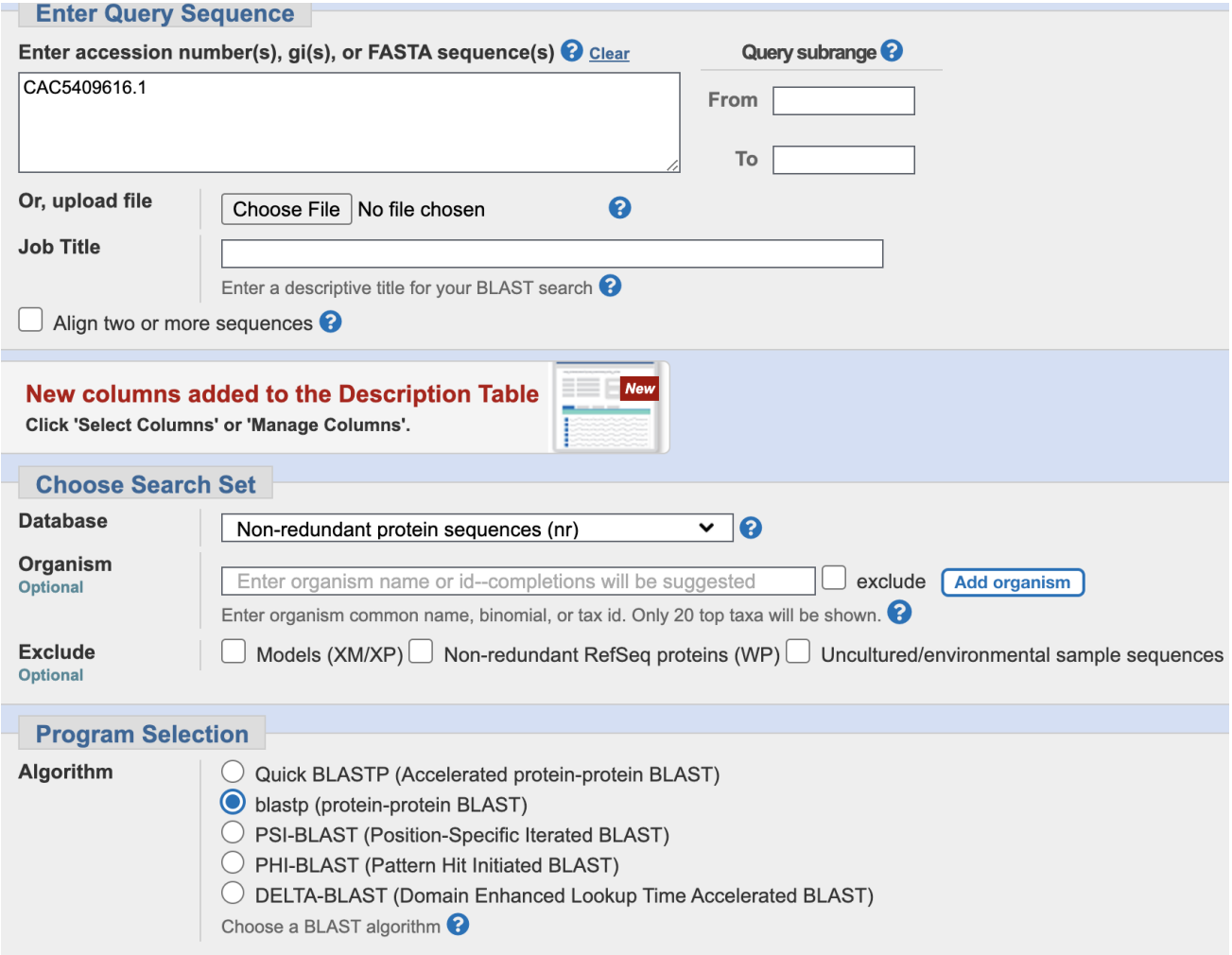
188 aa protein

Accession: CAC5409616.1 GI: 1866842361

BioProject Nucleotide Taxonomy

GenPept Identical Proteins FASTA Graphics

Blasting Options



Blasting Options

— Algorithm parameters

General Parameters

Max target sequences	100 Select the maximum number of aligned sequences to display
Short queries	☑ Automatically adjust parameters for short input sequences
Expect threshold	0.05
Word size	6
Max matches in a query range	0

Scoring Parameters

Matrix	BLOSUM62
Gap Costs	Existence: 11 Extension: 1
Compositional adjustments	Conditional compositional score matrix adjustment

Filters and Masking

Filter	☐ Low complexity regions
Mask	☐ Mask for lookup table only ☐ Mask lower case letters

BLAST Search **database nr** using **Blastp (protein-protein BLAST)**

Scores

- **Max Score**
  - The score of the best matching segment for local alignment, not global
- **Total Score**
  - The total scores of all matching segments found (same as max score if there is only one matching segment)
- **Query Coverage**
  - The percentage of the query sequence that aligned to some part of the match.
- **E-Value**
  - A statistical measure evaluating how likely it is that a match this good could occur by chance. Lower e-scores indicate that both sequences are truly similar and are not similar by chance alone. Identical sequences have e-scores of zero.
- **Max Indent**
  - The percentage of nucleotides that are identical between the query and the target sequences within the matching regions.

Blasting Results

GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	SIRT1 [Mytilus coruscus]	Mytilus c…	382	382	100%	9e-134	100.00%	188	CAC5409616.1
☑	NAD-dependent deacetylase sirtuin 1 [Mytilus galloprovincialis]	Mytilus g…	323	323	89%	1e-102	93.45%	826	VDI49146.1
☑	NAD-dependent protein deacetylase sirtuin-1-like isoform X3…	Mizuhop…	178	178	71%	3e-48	60.14%	869	XP_021368443.1
☑	LOW QUALITY PROTEIN: NAD-dependent protein deacetyl…	Pecten m…	178	178	63%	4e-48	65.29%	834	XP_033738334.1
☑	NAD-dependent protein deacetylase sirtuin-1 [Mizuhopecten …]	Mizuhop…	176	176	71%	2e-47	60.42%	850	OWF43165.1
☑	NAD-dependent protein deacetylase sirtuin-1-like [Pomacea …]	Pomacea…	165	165	60%	1e-43	66.37%	848	XP_025089963.1
☑	NAD-dependent protein deacetylase sirtuin-1-like isoform X2…	Mizuhop…	162	162	57%	9e-43	67.89%	749	XP_021368442.1
☑	NAD-dependent protein deacetylase sirtuin-1-like isoform X1…	Mizuhop…	162	162	57%	1e-42	67.89%	765	XP_021368441.1
☑	NAD-dependent protein deacetylase sirtuin-1 isoform X1 [Lin…	Lingula a…	157	157	60%	5e-41	60.53%	737	XP_013411402.1
☑	NAD-dependent protein deacetylase sirtuin-1 isoform X2 [Lin…	Lingula a…	157	157	60%	5e-41	60.53%	736	XP_013411403.1
☑	hypothetical protein Cfor_04474 [Coptotermes formosanus]	Coptoter…	155	155	76%	4e-40	52.78%	880	GFG40552.1
☑	unnamed protein product [Timema poppensis]	Timema …	149	149	59%	5e-40	60.71%	377	CAD7411610.1
☑	hypothetical protein C0J52_01051 [Blattella germanica]	Blattella …	155	155	71%	7e-40	53.62%	866	PSN54664.1

Blasting Results

Download GenPept Graphics ▼ Next ▲ Previous

SIRT1 [Mytilus coruscus]

Sequence ID: CAC5409616.1 **Length:** 188 **Number of Matches:** 1

Range 1: 1 to 188 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
382 bits(980)	9e-134	Compositional matrix adjust.	188/188(100%)	188/188(100%)	0/188(0%)

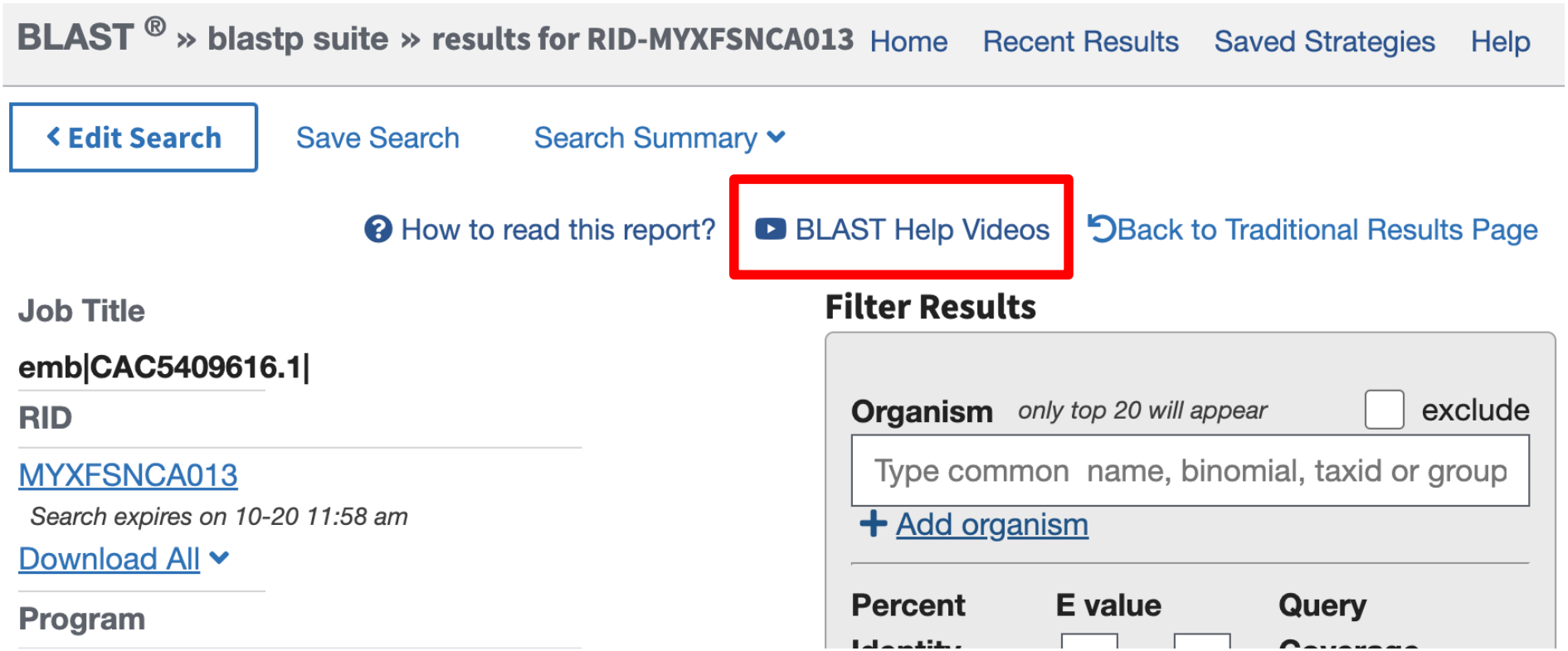
```
Query  1    MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ
            MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ
Sbjct  1    MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ

Query  61   IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT
            IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT
Sbjct  61   IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT

Query  121  DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ
            DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ
Sbjct  121  DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ

Query  181  MESMLALL  188
            MESMLALL
Sbjct  181  MESMLALL  188
```

Tutorials!


BLAST ® » blastp suite » results for RID-MYXFSNCA013
Home
Recent Results
Saved Strategies
Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title
emb|CAC5409616.1|
RID
MYXFSNCA013
Search expires on 10-20 11:58 am
Download All
Program
Filter Results
Organism
only top 20 will appear
exclude
Type common name, binomial, taxid or group
Add organism
Percent
E value
Query

Group Work

Playlist
https://www.youtube.com/playlist?list=PL7dF9e2qSW0azL2xOKAtxDW7QI8UU4XZ6

NCBI Minute: Enhancements to BLAST and Primer-BLAST
https://www.youtube.com/watch?v=LnkNhyTz4lo

NCBI Minute: Updated BLAST rRNA Databases for Identification
https://www.youtube.com/watch?v=I45-mmXM84U

NCBI Minute: QuickBLASTP - Rapidly Find High-scoring Protein Matches
https://www.youtube.com/watch?v=pO_a4e7QGRk

NCBI Minute: Using organism (taxonomic) information with standalone BLAST
https://www.youtube.com/watch?v=c-pFrvX5Aiw

Group Work!

GitHub Classroom working repository:
https://classroom.github.com/a/8ORH9mj-

Blast Tutorials!

- **Watch a tutorial**
- **You and your group to present one or two MAIN features**

- **Present on Friday:**
  - **One or two main features**
  - **Live demonstration**

STOP!
STOP!

Have a group member click the link, give your group a name and then have each person join that group.

Due on Friday

21 Oct 2022